COVID-19 Case Surveillance Public Use Data with Geography Utility Summary

Users should consider the level of completeness, including suppression levels when planning their analyses and use of public datasets. Privacy protections will suppress field values to reduce reidentification risks. Completeness varies by jurisdiction (i.e., state, local, and territorial) and time period. Variables are consistently coded to the value "Unknown" when jurisdictions specify in the case data submitted to CDC that the value is unknown, the value "Missing" when jurisdictions do not provide a value, and the value "NA" when the value is suppressed as part of privacy protections.

Dataset version: 2021-05-24

Total records in dataset: 25,607,582

Quick Summary

	all_fields	quasi_fields
total_fields	19	8
total_records	25,607,582	25,607,582
total_cells	486,544,058	204,860,656
missing_fields	103,438,730	13,361,352
missing_pct	21%	7%
complete_fields	383,105,328	191,499,304
complete_pct	79%	93%
unknown_fields	24,108,287	12,392,874
unknown_pct	5%	6%
suppressed_fields	15,514,873	13,702,737
suppressed_pct	3%	7%
available_fields	343,482,168	165,403,693
available_pct	71%	81%

Field-level Utility Summary

	suppressed	suppressed_percent	missing	missing_percent
case_month	0	0.0%	0	0.0%
res_state	738	0.0%	0	0.0%
res_county	1,811,398	7.1%	0	0.0%
age_group	286,822	1.1%	162,266	0.6%
sex	753,153	2.9%	33,217	0.1%
race	4,388,557	17.1%	2,073,839	8.1%
ethnicity	5,256,185	20.5%	1,678,851	6.6%
death_yn	1,205,884	4.7%	9,413,179	36.8%
records_with_any_field	7,458,274	29.1%	10,565,110	41.3%